CSC411 Machine Learning
Project 3: Fake News

Ariel Kelman
Student No: 1000561368

Gideon Blinick
Student No: 999763000

18 March 2018

1 Dataset Description

Describe the datasets. You will be predicting whether a headline is real or fake news from words that appear in the headline.
Is that feasible? Give 3 examples of specific keywords that may be useful, together with statistics on how often they appear in real and fake headlines.

Both datasets (real and fake) contain headlines about U.S. President Donald Trump. From the headlines, it is very much possible to determine with significant accuracy the percentage chance that a given headline is fake or real based on the precense of absence of certain words in the headline. For example, after performing some preliminary analyses on the headlines, we discovered that the following 3 words might be of particular use.

1. The word ”donald” appears 42.05% of the time in real headlines while appearing in only 17.55% of fake headlines. This 24.5% difference was by far the largest of any word.

2. The word ”the” appears 27.9% of the time in fake headlines while appearing in only 7.9% of real headlines for a difference of 20.02%.

3. The word ”trumps” appears 11.1% of the time in real headlines while appearing in only 0.3% of fake headlines for a difference of 10.8%.

2 Naive Bayes Implementation

3 Predictive Factors

4 Logistic Regression

5 Logistic Regression vs. Naive Bayes

6 Analysis of Logistic Regression

7 Decision Tree

7.1 Classification

Using the `sklearn` implementation of decision trees, we trained several decision trees to differentiate between the fake and real headlines. After some experimentation with the many parameters in the `sklearn DecisionTreeClassifier` (particularly with the maximum number of features used when looking for the best split), the default parameters gave the best results. Many of the setting served as ways to limit the size of the decision tree, so this result is not unexpected.

Decision trees with a maximum depth of $\{2, 3, 5, 10, 15, 20, 35, 50, 75, 100, None\}$ were built, and the results on the training, validation, and testing sets are shown in the figure below. The final point, not plotted, with no limit on the maximum depth, gave an accuracy of $1, 0.760.76$ on the training, validation, and testing sets respectively. As can be seen from the figure, the larger the depth of the decision tree, the greater accuracy achieved on the testing set. However, improvement is negligible on the validation and testing sets after a depth of 20. Without any other constraints (such as a minimum number of samples to split a node), a decision tree can reach perfect accuracy on the training set, as demonstrated above. Predictably, this leads to severe overfitting; showing the importance of using a validation and testing set to measure preformance.

Figure 1: Plot showing the preformance of `sklearn` decision trees with varying depth.

7.2 Visualization

The following image shows the first few layers of the decision tree with depth 20. It was generated by saving a text representation of the tree (as a `.dot` file), which was then visualized by using the

Figure 2: The top few layers of the decision tree (depth = 20).

It is interesting to look at the words being split on in these layers; X was a list of all words that appeared in either the fake or real datasets. These words are ‘donald’ (23), ‘trumps’ (1604; note the double appearance), ‘the’ (81), ‘hillary’ (44), and ‘trump’ (3). The exact same process of training was run ignoring stop words, but results were slightly worse; graphs similar to those mentioned above can be found in the `resources` directory (each filename indicates whether stop words were included).

Note that as one moves down the tree, the entropy reduces, which is to be expected - the words that give in the `resources` directory; the text represenations of many trees that were generated during training are saved in `resources/part7`.

7.3 Comparison of All 3 Classifiers

All three classifiers preform much better than random guessing...

8 Information Theory